EXHIBIT OO

From:	Joshua Stein
To:	Lauter, Judd; Dunning, Angela L.; Kathleen Hartnett; mlemley@lex-lumina.com; Ghajar, Bobby A.; Ghazarian, Colette A
Cc:	Jesse Panuccio; Poppell, Cole A; Biksa, Liene; Weinstein, Mark; Stameshkin, Liz; Alvarez, Jessica; Holden Benon; Christopher Young; Aaron Cera; Cadio Zirpoli; Joe Saveri; Margaux Poueymirou; Ashleigh Jensen; Rya Fishman; Matthew Butterick; Nada Djordjevic; James Ulwick; Bryan L. Clobes; Mohammed Rathur; Amy Keller; David Straite; Ruby Ponce; Alexander Sweatman; Heaven Haile; Llama BSF; Josh Schiller; David Boies; Maxwell Pritt; z/Meta-Kadrey
Subject:	Letter to Meta re Additional Discovery Issues
Date:	Wednesday, October 9, 2024 5:50:54 AM
Attachments:	10092024 Letter to Meta re Additional Discovery Issues.pdf

Counsel,

Please see the attached correspondence.

Best,

Josh

Joshua Michelangelo Stein
Partner

BOIES SCHILLER FLEXNER LLP
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(t) +1 415 293 6813
(m) +1 617 365 3991
jstein@bsfllp.com
www.bsfllp.com

BSF | BOIES SCHILLER FLEXNER

October 9, 2024

SENT VIA EMAIL

Maxwell V. Pritt
Boies Schiller Flexner LLP
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
mpritt@bsfllp.com
(415) 293-6800

Bobby Ghajar
bghajar@cooley.com
Colette Ghazarian
cghazarian@cooley.com
Cooley LLP
133 2nd Street, Suite 400
Santa Monica, California 90401

Angela L. Dunning
adunning@cooley.com
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304

Mark Weinstein
mweinstein@cooley.com
Kathleen Hartnett
khartnett@cooley.com
Judd Lauter
jlauter@cooley.com
Elizabeth L. Stameshkin
estameshkin@cooley.com
Cooley LLP
3175 Hanover Street
Palo Alto, CA 94304-1130

Re: *Kadrey v. Meta*, Case No. 3:23-cv-03417-VC, Additional Issues Concerning Meta's Discovery Responses and Productions

Dear Counsel:

We write concerning several deficiencies in Meta's objections and responses to Plaintiffs' discovery requests and Meta's document productions.

In the interest of resolving these issues without court intervention, we identify the specific discovery deficiencies below and request that Meta confirm by 5:00 p.m. pacific on Thursday, October 10, whether it will supplement any of its responses and productions

BSF

as requested herein. If so, please identify which ones. For any deficiency that Meta declines to correct, we request a meet and confer to address those matters on Friday, October 11, otherwise we will raise these issues with Judge Hixson in accordance with his Discovery Standing Order.

I. Requests for Production

It appears that Meta failed to search all relevant document repositories for documents responsive to at least RFP Nos. 1-12 and 36-38, including Meta Manifold, RSC, and GTT. Meta_Kadrey_00065246, for example, shows that in addition to S3, Meta stored training datasets in Hive and locations called "RSC" and "GTT." And Meta_Kadrey_00065314.00011 refers to a storage location called "Meta Manifold." Yet not one of these non-custodial data sources are in the list of ESI sources Meta claims it searched, suggesting that additional responsive documents and training data are missing. Accordingly, Plaintiffs request that Meta search for and produce all relevant and responsive non-privileged documents and/or data in these non-custodial data sources (and all other potentially relevant non-custodial data sources that Meta has not identified to Plaintiffs), including all training data. In addition, Plaintiffs request that Meta produce all non-privileged documents responsive to RFP Nos. 36 and 38, not just those "sufficient to show."

In RFP No. 49, Plaintiffs requested "All Documents and Communications Concerning the decision to release the Meta Language Models under what Meta calls an 'open source' license" on December 27, 2023. On February 23, 2024, Meta responded that it would search for and produce only documents "sufficient to show the reasoning behind Meta's decision to make its Meta Language Models (as construed above) available to the public under an open license." We do not agree with Meta's limitations. Please confirm that Meta will search for and produce *all* non-privileged, responsive documents concerning *all* Llama models immediately, not simply what is "sufficient to show" Meta's reasoning.

Meta refused to produce documents responsive to RFP No. 54. That RFP, which Plaintiffs served on March 20, 2024, seeks "All Documents and Communications Concerning any decision by You to not develop an interface for end users to interact with any of the Meta Language Models." On April 19, 2024, Meta objected "to this Request as overbroad, unduly burdensome, and disproportionate to the needs of the case." This objection is not well founded. The RFP is plainly relevant to this case, including because it pertains to the use of Meta's models, a significant factor in the fair use analysis. Please confirm that Meta will search for and produce all non-privileged documents responsive to RFP No. 54.

Meta also refused to produce documents responsive to RFP No. 59. That RFP, also served on March 20, 2024, seeks "Documents and Communications Concerning the

BSF

ability of any Meta Language Model to output fictional works." On April 19, 2024, Meta objected "to this Request as vague and ambiguous as to the phrase 'fictional works,'" among other things. Again, Meta's objections are not well founded—this RFP also seeks documents relevant to, among other things, fair-use factor one, as evidenced by Meta's response to ROG No. 1, Set 2 (further discussed below). Please confirm that Meta will search for and produce all non-privileged documents responsive to RFP No. 59.

Finally, Meta refused to produce documents responsive to RFP Nos. 22–28, primarily on grounds of burden and relevance, including a claim that others would be more "directly involved" in training data. But these RFPs seek documents that would speak directly to the state of mind of Meta's leadership, which is relevant both to willfulness and fair-use factor one. Please confirm that Meta will search for and produce all non-privileged documents responsive to RFP Nos. 22–28.

II. Interrogatories

In response to ROG No. 3, Meta stated that it would identify documents or information sufficient to show "steps undertaken as a part of RLHF of Llama 2, if any, that reduce the likelihood that the model could reproduce verbatim content from any training data." However, Meta's supplemental ROG response does not provide such information. Please confirm that Meta will supplement its response to ROG 3 to provide this information.

In its response to ROG No. 4, Meta responded only to part (b) of this Request, identifying certain individuals involved in assessing the risk, safety, and alignment of Meta's Llama models. Meta failed to respond to parts (a), (c), (d), and (e) of this Request. Please confirm that Meta will supplement its response to ROG 4 and respond to the other parts of this Request.

In response to ROG No. 5, which asks that Meta identify any and all "agreements" relating to the data used to train its Llama models, Meta limited the Request to "written, signed agreements." We do not agree with that limitation. Meta should identify any type of agreement, even if informal, oral, unexecuted, or in progress. Please confirm that Meta will supplement its response and respond to ROG No. 5 as written and identify all agreements relating to the data used to train Meta's Llama models, even if informal, oral, unexecuted, or in progress.

Meta refused to respond to ROG Nos. 13, 14, and 15 on the basis that some of Plaintiffs' earlier interrogatories contained subparts and thus counted towards the 25-interrogatory limit under Rule 33(a)(1). As you know, the rule regarding impermissible subparts in interrogatories applies only to discrete subparts that ask about *separate and*

BSF

distinct subjects; subparts count as one interrogatory so long as they are logically or factually related. *Synopsys, Inc. v. ATopTech, Inc*, 319 F.R.D. 293, 297 (N.D. Cal. 2016).

Meta's refusal to answer ROG Nos. 13, 14, and 15 on this ground is also inconsistent with Meta's subsequent substantive responses to Plaintiffs' second set of interrogatories. Please confirm that Meta will supplement its responses and respond substantively to ROG Nos. 13, 14, and 15.

Finally, Plaintiffs request that Meta agree that the parties may serve an additional 23 interrogatories. Absent such agreement, Plaintiffs will raise this request with Judge Hixson.

III. Incomplete or Mangled Data

Meta's employees utilize a platform known as facebook.workplace.com for internal communications, which is included in Meta's list of ESI locations (*see* September 19, 2024 6:01 pm email RE: Kadrey v. Meta: Disclosure of Data Sources from Liz Stameshkin to Holden Benon, et al."). However, the communications that Meta produced from Workplace have been converted into a mangled format that does not resemble the original messages. For instance, the document Meta_Kadrey_00074729.jpg illustrates how a Workplace Chat has been transformed into plain text when emailed to a custodian. This conversion process results in a loss of formatting, hyperlinks, emojis, images, and other integral elements present in the original communications. Meta's production of these documents therefore is deficient because they were not provided in their native format, resulting in the stripping away of hyperlinks, comments, and other essential features. Please confirm that Meta will re-produce these communications in a non-transformed manner.

Additionally, it has become evident that Meta failed to search for and produce all responsive documents from relevant custodial data sources. Custodial documents from messaging platforms including, but not limited to WhatsApp, SMS, Discord, Signal/Telegram, Airstore data loader (*see* Meta_Kadrey_00033932), and iMessage have largely not been produced; indeed, Meta did not produce any WhatsApp messages until the eve and morning of Mr. Al-Dahle's deposition. Yet it appears that Meta AI was accessible through both WhatsApp and Messenger. *See* Yann LeCun (@ylecun), X (Jan. 18, 2024, 10:49 PM) https://twitter.com/ylecun/status/1748236213784281154; Meta_Kadrey_00032996. It also appears that Meta employees used iMessages to communicate about issues relevant to this case—*see, e.g.*, Meta_Kadrey_00089014–17—yet few such messages have been produced. Notably, one message discusses "how Open AI was interpreting fair use, which set a market precedent for how the industry could approach fair use of copyrighted material." Meta_Kadrey_00089015. Please confirm

BSF

that Meta will search for and produce all non-privileged responsive documents on the messaging platforms identified above, and any other potentially relevant custodial data sources.

IV. Additional Custodians

Plaintiffs request that Meta add the following individuals as custodians: Luke Zettlemoyer, Armand Joulin, Steven Roller, Sean Bell, Kenneth Heafield, Brian Gamido, Elisa Garcia Anzano, Xavier Martinet, Guillame Lample, Ragavan Srinivasan, Moya Chen, Santosh Janardhan, Arun Rao, Michael Mayer, and Jennifer Pak.

- Mr. Zettlemoyer is a research director at FAIR (Fundamental AI Research). Kambadur Dep. at 282:21–282:22. As a director at FAIR, he has exclusive knowledge regarding the development of Llama and the data used to train it. He may also have been consulted about Genesis. *Id.* at 38:15–38:18.

- Mr. Joulin is one of the managers responsible for the development of Llama 1 series of models. Kambadur Dep. at 44:3-44:6. He has stated: "I see two problems, however, with what you are describing. One, we use data that contains copyrighted information; e.g., and Common Crawl." *Id.* at 154:6–154:11. Mr. Joulin has knowledge concerning the training sets and/or data used to create Llama 1.

- Steven Roller is a former research engineer on Ms. Kambadur's team. Kambadur Dep. at 168:6–168:7.

- Sean Bell is a manager supporting Meta's data foundations team in generative AI. Kambadur Dep. at 195:16–195:18. Mr. Bell currently is the head of Llama Data and likely has knowledge concerning the data used to create the Llama models.

- Kenneth Heafield is the current manager of Todor Mihaylov. Mihaylov Dep. at 103:20–103:22. Mr. Heafield tried to get Internet Archive as a potential data source for Llama, but Mr. Mihaylov did not know if he succeeded. *Id.* Mr. Mihaylov also said Mr. Heafield may have mentioned that Meta was considering paying for the data they use. *Id.* at 141:1–141:15.

- Brian Gamido is part of the AI Business Development team. Mr. Gamido was assigned the task of reaching out to content owners to discuss potential licensing opportunities. Meta_Kadrey_00045315-23.

BSF

- Elisa Garcia Anzano was part of the AI Business Development team. Ms. Garcia Anzano was assigned the task of reaching out to content owners to discuss potential licensing opportunities. Meta_Kadrey_00045315-23.

- Xavier Martinet is a research engineer at FAIR (Fundamental AI Research) who is familiar with Meta's decision to use pirated materials rather than paying for licenses. Meta_Kadrey_00074729.

- Guillame Lample was a research scientist at FAIR with knowledge of the books in Meta's training dataset.

- Ragavan Srinivasan is the head of product at Facebook AI.

- Moya Chen is a former research engineer at FAIR who has knowledge on the downloading and processing of LibGen1 and LibGen2.

- Santosh Janardhan is the head of infrastructure at Meta. Meta_Kadrey_00046268 shows AI infra was led by Mr. Janardhan.

- Alexis Björlin is a former vice president of infrastructure at Meta. Meta_Kadrey_00046268 shows AI infra training was led by Mr. Björlin.

- Arun Rao is the lead product manager for Generative AI at Meta whose name appears on key documents regarding product uses.

- Michel Meyer is the group product manager working on Core Learning and Reasoning within FAIR.

- Jennifer Pak is Lead Counsel on intellectual property risks in Generative AI at Meta.

V. ESI Protocol

As discussed in Plaintiffs' accompanying discovery meet and confer letter, Plaintiffs request that Meta immediately provide its lists of search terms and hit counts in accordance with ESI Order. *See* ECF No. 101 at 4 ("The Producing Party will then apply the original and new search terms to the targeted document set and disclose the original and new terms to the Requesting Party, along with a hit report.").

In addition, Plaintiffs request that Meta identify all potentially relevant non-custodial and custodial data sources in its possession, custody, or control, and identify all

potentially relevant witnesses employed by Meta at any point during the relevant time period (i.e., the class period). We recognize the ESI Order currently does not require exchanging that information, and we are willing to amend the ESI Order if Meta prefers doing so. Please confirm that Meta will produce this information, whether under an amended ESI Order or otherwise. Absent agreement to do so, we will raise this issue with Judge Hixson.

VI. Relevant Time Period

In its responses and objections (and its productions), Meta unilaterally and improperly limited the relevant period to January 1, 2022, to the present. This artificial limitation is grossly inappropriate because the proposed class period begins on July 7, 2020. Please immediately supplement responses and productions to remove this objection and respond to Plaintiffs' requests based on the entire class period, including producing all responsive documents for that period.

VII. Fair Use (ROG No. 1, Set 2)

On September 30, 2024, Meta responded to Plaintiffs' ROG No. 1, Set 2, stating that it "intends to rely on . . . documents produced in this litigation." To the extent that Meta is aware of the documents produced in this litigation that it intends to rely on, those documents should be identified, and Meta must provide a basis for believing such documents support its view. Please confirm that Meta will amend its response to this ROG and provide this information.

Moreover, throughout this litigation, Meta has consistently claimed that any discovery regarding Meta's efforts to prevent its Llama Models from outputting training data comprising copyrighted material is irrelevant given the dismissal of Plaintiffs' output claims. However, Meta's ROG response makes clear that Meta intends to rely on its "efforts to minimize the models' ability to memorize and/or output training data verbatim" in its fair use-factor one defense. Please confirm that Meta will update its discovery productions and responses to include all documents and communications regarding its efforts both to minimize "memorization" and to prevent its Llama Models from outputting copyrighted material on which its Llama Models have been trained.

VIII. Advice of Counsel (ROG No. 2, Set 2)

On September 30, 2024, Meta responded to Plaintiffs' ROG No. 2, Set 2, that it "does not presently intend to assert the advice of counsel defense." Meta's ambiguous

BSF

qualification ("presently intend") is improper and insufficient. Please confirm that Meta will amend its response to this ROG and answer the questions without equivocation, either that Meta will assert the defense or that it will not do so.

* * *

Please let me know when you are available on Friday, October 11 to meet and confer regarding the issues identified above.

Plaintiffs reserve all rights, and the exclusion of any issues regarding Meta's discovery responses and productions in this letter does not suggest and cannot imply that Plaintiffs waive any rights thereby, including with respect to any issues already subject to meet and confer discussions.

Sincerely,

/s/ Maxwell V. Pritt
Maxwell V. Pritt
Boies Schiller Flexner LLP